Document made available under the Patent Cooperation Treaty (PCT)

International application number: PCT/DK04/000914

International filing date: 23 December 2004 (23.12.2004)

Document type: Certified copy of priority document

Document details: Country/Office: DK
Number: PA 2003 01940
Filing date: 27 December 2003 (27.12.2003)

Date of receipt at the International Bureau: 11 February 2005 (11.02.2005)

Remark: Priority document submitted or transmitted to the International Bureau in compliance with Rule 17.1(a) or (b)

Kongeriget Danmark

Patent application No.: PA 2003 01940

Date of filing: 27 December 2003

Applicant: (Name and address)
Torben F. Ørntoft
Helgesvej 19
DK-8230 Aabyhøj
Denmark

Title: Classification of Colon Cancer

IPC: -

This is to certify that the attached documents are exact copies of the above mentioned patent application as originally filed.

Patent- og Varemærkestyrelsen
Kongeriget Danmark

Patent- og Varemærkestyrelsen
Økonomi- og Erhvervsministeriet

03 February 2005

Susanne Morsing

PATENT- OG VAREMÆRKESTYRELSEN

Abstract

Precise classification of a tumor is imperative to apply the best possible therapy for the individual patient and to make prediction about clinically outcome. The standard treatment of patient with Dukes' C colon cancer includes post-chirurgic chemotherapy, whereas Dukes' B patients receive no chemotherapy. Patient with CRC tumors exhibiting microsatellite instability (MSI) have a better prognosis compared to patients with microsatellite stable tumors (MSS) but recent research showed that only patients with MSS benefit from chemotherapy. It is therefore of clinical relevance to identify patient with MSS tumors for chemotherapy independent on tumor stage.

The aim of this study was to build a robust classifier based on gene expression to separate MSS from MSI tumors. The robustness was achieved by collecting the tumors from 14 different clinics in two different countries, isolating RNA with two methods at three different sites, and labeling RNA in 10 separate batches. DNA microarray analysis was performed on 38 Danish and 64 Finnish tumors from primary CRC patients and 17 normal samples. Unsupervised hierarchical clustering analysis identified microsatellite instability as the main clinical features separating the samples into groups. In addition we found a weaker but clear separation as to the country of origin of the samples. Permutation analyses demonstrated that both the microsatellite-instability status as also the country of origin of the samples were highly significant signatures in the expression data. Removal of country specific genes improved the separation of MSS from MSI in unsupervised classification as demonstrated by multidimensional scaling.

Supervised classification of the 102 tumor samples using a maximum likelihood classifier with a crossvalidation loop resulted in 7 MSI samples being classified as MSS and 1 MSS sample being classified as MSI. Re-evaluation including IHC and specific genes expression levels of the misclassified MSI tumors indicated that 6 of these tumors were probably truly MSS One MSI tumor was a signet ring cell carcinoma with a low tumors fraction Based on this, we excluded the

misclassified tumors and re-build the classifier and tested its performance using from 10 or 100 genes All 94 tumors samples were classified correctly into MSS and MSI.

Classifiers should be if tested with an independent testset to prove its strength and more robustness. The large difference between MSS and MSI tumors and the large number of tumors allow us to separate our dataset into a set for training the classifier and an independent set for testing the classifier. We selected 25 MSI and 30 MSS samples for selecting of optimal classification genes and used the remaining tumors as an independent test sets for evaluation of classification performance of these genes. Since the performance of a classifier may depend on the tumors dedicated to the training set we tested the classifier by permutation analysis. The final classifier was based on 8 genes and classified MSS and MSI tumors with 98.2% precision. The MSS tumors were identified with a sensitivity of 99.8% and with a specificity of 93.8%.

Classification of Microsatellite Instable Colorectal Cancer

Introduction

Colorectal is the fourth most frequently diagnosed malignancy and the second most common cause of cancer death in the western world. Extensive investigated within the past decade has point towards two alternative genetic pathways in the development of cancer, the mutator phenotype featuring tumors with microsatellite instability (MSI) and the suppressor pathway represented by chromosomal instable but microsatellite stable (MSS) tumors The majority of the most common hereditary CRC syndrome HNPCC belongs to the group of MSI tumors.

MSI has been defined as a change of any length due to either insertions or deletions of repeating units in a microsatellite within a tumor compared to normal tissue and is caused by an underlying defect in the mismatch repair (MMR) system. (Boland et al, CR 1998, 58:5248). A compromised MMR system commonly affect genes that include or are linked to microsatellite repeat regions such as TGFβRII, ILGF, E2F-4 and BAX (Markowitz et al 1995), genes that are rarely mutated in MSS tumors Furthermore, MSI tumors are diploid and show no loss of heterozygosity whereas MSS tumors demonstrate a wide variety in chromosomal number and extensive LOH

The MSI pathway may either be sporadic or hereditary (HNPCC) and whereas the disruption of the MMR system in sporadic MSI tumors is most often caused by somatic methylation of the MLH1 promoter more that 90% of HNPCC cancers are caused by germline mutations in MLH1 or MSH2.

The MSS pathway to cancer begins with the inactivation of tumor suppressor genes, such as APC/β-catenin genes, followed by activation of oncogenes and inactivation of additional tumor suppressor genes, commonly with a high frequency of allelic losses and cytogenetic abnormalities and abnormal DNA tumor content.

Crude survival data suggest that patients with HNPCC have a better prognosis than those with sporadic disease and studies have also shown that MSI is an independent indicator of good prognosis A large recent study shown that MSS benefit from 5-FU treatment/leucovin treatment (Ribic et al., 2003) in contrast to MSI cancer patients gained no advantage in survival This is the exact opposite conclusion from earlier studies, which however used probe collection.

The recognition of different homogeneous groups of CRC with different pharmacological profiles is mandatory for designing the best individual therapy for the individual patient. Many studies have defined the pathoclinical trait of MSI and MSS tumors. MSI positive cancers most frequently found in the right side of the colon, they tend to be of less differentiated, they tend to be larger in size, are often mucinous and often exhibit extensive infiltration by lymphocytes.

Other studies have addressed the classification tumors either into dukes' stages B and C (Frederiksen and Orntoft, 2003) or different levels of microsatellite instability (Mori et al . 2003)

Computational scientists in collaboration with medical scientist today readily download datasets from the Internet and combine these even across different platforms. It is well know that noise and disparities in experimental protocols set strong limits to this form of data integration. Platform biases may originate from different probes, i.e. cDNA probes versus oligonucleotides, labelling procedures, quality etc. But even studies conducted in one laboratory using a single platform underlie the risk of serious biases. Often the samples used have been collected in different clinics with adverse procedures The time from resection of a tumor to preservation can change the expression of a number of genes as response to ischemia (Huang et al . 2001), the amount of normal tissue in the tumor may be very different, information on the location and type of the tumor may not be available. This may lead to more or less systematic errors like e.g. samples clustering according to batch of labeling (Mori et al . 2003) or procedure for sampling or trimming of the tumor tissue.

Materials and Methods

Biological material From the Danish and Finnish CRC tissue banks 102 primary colorectal cancers and 17 macroscopically normal colon epithelium samples from the oral resection edge were chosen. Only adenocarcinomas from Dukes' stage B and C were included, however, these represented a broad spectrum of tumors in relation to location, heredity, microsatellite instability status, and origin of the patient. All tumors were collected in the period from 1994 to 2002. 75 tumor samples were collected at nine different clinics in Finland and 47 samples were collected at four different clinics in Denmark, 37 were Dukes' B, 65 Dukes' C, 25 were sporadic microsatellite highly instable (MSI-H), 17 HNPCC and MSI-H, and 59 were sporadic microsatellite stable (MSS) (table 1) None of the patients received pre-operative radiation or chemotherapy

Microsatellite analysis. From all tumor samples available as paraffin blocks, ten sections were cut at 10µm and stained with haematoxylin. The first and last section was cut at 4 µm, stained with haematoxylin, and routinely mounted. These two sections were used for the identification of tumor and normals cells from each sample. Regions enriched in tumor cells (more than 90%) were microdissected from these sections and DNA was extracted using a Puregene DNA extraction kit (Gentra Systems, Minneapolis, MN) DNA from blood samples was used as control when available, otherwise normal tissue was microdissected from the tissue sections The samples were analyzed for microsatellite instability according to the NCI guidelines (Boland et al) using markers BAT25 and BAT26 as previously described (Loukola et al. 2001). Some of the Danish samples were difficult no definitive result could be obtained.

RNA purification Colorectal specimens were obtained fresh from surgery and were immediately snap frozen in liquid nitrogen either as was, in OCD-compound or in an SDS/guadinium thiocyanate solution. Total RNA was isolated using RNAzol (WAK-Chemie Medical) or spin column technology (Sigma) according to the manufacturer's instructions.

Preparation of labelled aRNA target Ten µg of total RNA was used as starting material for the target preparation as described (Dyrskødt et al., 2003). Briefly, the first and second strand cDNA synthesis was performed using the SuperScript II System (Invitrogen) according to the manufacturers' instructions except using an oligo-dT primer containing a T7 RNA polymerase promoter site Labelled aRNA was prepared using the BioArray High Yield RNA Transcript Labelling Kit (Enzo). Biotin labelled CTP and UTP (Enzo) were used in the reaction together with unlabeled NTP's. Following the IVT reaction, the unincorporated nucleotides were removed using RNeasy columns (Qiagen)

Array hybridization and scanning These procedures were performed at described in detail elsewhere (Dyrskødt et al, 2003) Briefly, 15 µg of cRNA was fragmented, loading onto the Affymetrix HG_U133A probe array cartridge and hybridized for 16 h. The probe arrays were then washed and stained in the Affymetrix Fluidics Station and scanned using a confocal laser-scanning microscope (Hewlett Packard GeneArray Scanner G2500A) The readings from the quantitative scanning were analyzed by the Affymetrix Gene Expression Analysis Software (MAS 5.0).

Data processing

The arrays were normalized using RMA (robust multi array, Irizarry et al., 2003) Redundancy of probesets as defined form Unigene build 168, was reduced by removing probesets with high correlation (>0.5) over all samples.

Unsupervised agglomerative hierarchical clustering

For hierarchical expression cluster analysis 1239 genes with a variation across all samples greater than 0.5 were median-centred and normalized to a magnitude of 1. Samples and genes were then clustered using average linkage clustering with a modified Person correlation as similarity metric (Eisen et al., 1998). The cluster dendrogram was visualized with TreeView (Eisen).

Group testing

We make a statistical test where the p-value is evaluated through permutations. For each group and gene we calculate the average and the sum of squared deviations from the average We then sum these over the genes and the groups:

$$S_1 = \sum_{\text{groups}} \sum_{\text{genes}} \left(X_{ij} - \bar{X}_{gr(i)j}\right)^2$$

This expression is calculated for joining DK with SF and MSI with MSS such that we end up with two groups. The sum of squared deviations is denoted S_2. As a test statistic we use S_1/S_2. A small value indicates that there is a real reduction in the deviations when going from 2 to 4 groups and thus the groups have a real significance. To judge if a value is significantly small we use permutations. For each of the four groups left when joining DK and SF we randomly allocate the

members to a pseudo DK and pseudo SF in such a way that the number of members in each group are as in the original data

To get an understanding of this separation we performed a test to see if this is caused by a few genes or if many genes are involved. For this test we calculated $S_1 = \sum_{genes} S_1(gene)$ and similarly with $S_2 = \sum_{genes} S_2(gene)$ For each gene j we used the test statistic $S_1(j)/S_2(j)$ (Table 2.2)

Multidimentional Scaling

Multidimentional scaling was performed in R and visualized in a two-dimentional plot.

Microsatellite status classifiers

Maximum likelihood classifiers were build as described in Dyrskødt et 2003 For details refer to the text.

Results

Hierarchical Clustering

The clinical specimens used in this study were collected in two different countries from 14 different clinics in the period 1994 to 2001. The samples were selected to keep a balanced representation of microsatellite instable (MSI) and microsatellite stable (MSS) tumors from both the right- and left-sided colon. The MSI class was represented both by sporadic MSI and hereditary MSI (HNPCC) tumors. Only Dukes' B and Dukes' C tumor samples were included (table 1). Before any attempt to divide a diverse sample collection into distinct classes we analyzed the data for systematic bias that may have been introduces during the experimental procedures. A fast and easy way to discover both true distinct classes as well as systematic biases in the data is to perform a hierarchical clustering.

The phylogenetic tree resulting from hierarchical clustering on 1239 genes (fig. 1) reveals that the main separating factor is microsatellite status. On the upper trunk we find two clusters represented mainly by normal biopsies (14/21) and MSS tumors (18/25), respectively. The lower trunk is divided into a MSI cluster (30/36) and a second MSS cluster (MSS2-cluster) (34/37). A closer inspection of the two MSS clusters unveil that one is dominated by Danish samples (19/25) and one by Finnish samples (26/37 check). Also, it is worth to notice that the MSI cluster contains a vast majority of Finnish samples (32/36) and that the sporadic MSI samples are interspersed among the hereditary samples. The normal biopsies cluster tight together with a slight tendency to separation according to origin. Tree normal samples cluster within the MSI cluster indicating that these samples may have been resected to close to the tumor lesion.

Inspection of the gene cluster dendrogram shows that the two groups of MSS tumors are mainly separated by a cluster of approximately 150 genes being upregulated in the Danish samples (data not shown) indicating that there is truly a systematic difference between Danish and Finnish samples.

Difference between Danish and Finnish tumor samples

Based on these observations and concentrating on the tumor samples, we excluded normal samples and formed the following four virtual tumor groups: Danish MSI (MSI-DK), Danish MSS (MSS-DK), Finnish MSI (MSI-SF) and Finnish MSS (MSS-SF). Using 5082 genes with a variance above 0.2, we tested if all the groups are significant or if some of the groups can be joined. We considered the two possibilities of joining DK and SF, and joining MSI and MSS and made a statistical test where the p-value is evaluated through permutations (Table 2). We see that our test value S1/S2 is smaller in our groups than in all permutations demonstrating a very clear separation between DK and SF and also a very clear separation between MSI and MSS. To get an understanding of this

separation we performed a similar test to see if this is caused by a few genes or if many genes are involved For both the DK-SF and MSI-MSS, we observe that many genes cause this effect (Table 3).

When a property is present that influences a large proportion of the genes and if this influence can vary from sample to sample the ordinary normalization procedures can give misleading results. To analyze this we calculated distances by multidimensional scaling between samples with and without re-scaling of the data and plotted these in a two-dimensional plot (Fig 2 a, b). We find that re-scaling of the data improves the separation of the groups significantly. Next, we identified and excluded 816 genes that separate DK from SF with a t-value numerically greater than 2, which lead to a further improvements (Figure 2c). (This plot is not entirely unsupervised since the groups have been used to remove gene). We now see a separation of MSI and MSS with Danish and Finnish cases mixed The MSI-DK samples are not completely separated as they are found both between the MSI-SF and the MSS samples. At this point we looked at the identity of the genes that were responsible for the separation of DK from SF. The two genes with the highest fold change were S100A8 (4 3 fold upregulated) and Hemoglobin B (5.6 fold upregulated). These genes were also two of the most prominent genes identified to be upregulated in tumors as a function of time before the samples were frozen after resection (Huang et al , 2001: Yeatman personal communication) Thus these genes may represent a response to ischemia and indicate that the sample procedures in Denmark and Finland differed in a systematic way.

Construction of an MSI-MSS classifier Next, we build a maximum likelihood classifier with a 'leave one out' crossvalidation scheme to classify MSI and MSS tumors In order to evaluate the effect of systematic differences between DK and SF we constructed classifiers based on 24 genes with and without re-scaling of the data and with and without the 816 DK-SF classifier genes All

classifiers result in eight errors and these errors are in all cases the same tumors. The specific genes used for classification show a overlap of 14-18 of the 24 genes. We see that the classifier works well and seems to be independent of the removal of genes and rescaling, which is due to a large difference between MSI and MSS tumors

It is noteworthy that seven out of eight errors in the classification of 102 tumor samples were MSI samples being classified as MSS and that six of these were from Denmark. In order to understand this, we re-evaluated the clinical data for these tumors (table 2). One tumor turned out to be a signet cell carcinoma with a low ratio of tumor to normal cell, which may make a correct classification difficult. The remaining six MSI tumors were all left sided and are of high to middle grade of differentiation. All five of these six MSI-DK tumors that were stained for MLH1 and MSH2 in IHC were positive for both. The single MSS tumor that misclassified was right sided

We then looked at expression levels of a number of genes have been described in detail in relation to MSI tumors (Fig 3). Most of the misclassified MSI tumors had a class aberrant gene expression levels for most for the analyzed genes. Thus expression levels of MLH1, TGFβ induces protein (TGFBI) and cytokeratin (CK23) were higher compared to MSI whereas thymidylate synthase (TYMS) expression was lower. Based on these data, we conclude that the misclassified tumors probably are MSS but are showing an aberrant behavior in the microsatellite test

We now excluded the 816 ischemia genes and the eight outlier samples and rebuild our classifier. We decided to let the classifier select 10 or 100 genes and we choose those genes that were included in at least 70% of the crossvalidation loops This resulted in 8 and 96 genes, respectively (Table 4) and resulted in correct classification of all tumors.

Training of a classifier and subsequent testing with an independent dataset

It is commonly accepted that a classifier should be tested with an independent testset. The number of samples needed for training of a classifier is dependent of the difference between the groups to be classifier. Above we have demonstrated that difference between MSS and MSI is relatively large and can be well separated with only eight genes. After the exclusion of the misclassified tumors, we therefore selected 25 MSI and 30 MSS tumors randomly and allowed the classifier to choose 10 genes to be used for classification. The remaining 10 MSI and 29 MSS tumors were used as an independent test set. The performance of a classifier may be dependent on which samples are used for training and testing. Therefore we made 100 permutations of training and test sets and calculated the number of errors in crossvalidation of the training set and the number of errors in the classification of tumors in the test set (Table 5) In the 100 permutations of the training sets we find a mean of 7.2% errors of MSI (n=25, range 0-4) and 0.13% errors of MSS (n=30, range 0-1). Of the ten genes, eight genes were used in at least 70% of the crossvalidation loops and therefore used for classification of the test set (Table 5). Using these eight genes, the mean number of errors in the permutated test sets was 6.8% for MSI (n=10, range 0-3) and 0 17% for MSS (n=29, range 0-3) resulting in an overall performance of the classifier of 98.2% correct classification (n=39, range 0-3). In terms of sensitivity and specificity, the classification of MSS tumors was classified with a sensitivity of 98.2% and with a specificity of 96 2%.

Using the 8-gene classifier, classification of tumors consisting of 26 patients with Dukes B tumors showed 14 to be MSI and 12 to be MSS. The overall survival was highly significantly related to the classification as no individual died in the MSI group whereas 9 out of 12 died in the MSS group (Figure 4). Thus, the 8 gene classifier clearly proved to be a strong predictor of survival in Dukes B and it can be used to select patients who need adjuvant chemotherapy, namely those classified as MSS.

In the Dukes C group 47 were classified as MSS and 13 as MSI There was no significant difference in the survival between these groups. A trend was that the MSI showed a poorer survival than the MSS, contrary to Dukes B patients This difference can be attributed to the fact that a recent large study has shown that chemotherapy only benefit the MSS tumor patients, thus improving their survival to a level comparable to that which is characteristic of MSI tumor patients.

Agrawal D, Chen T, Irby R, Quackenbush J, Chambers AF, Szabo M, Cantor A, Coppola D, Yeatman TJ Osteopontin identified as lead marker of colon cancer progression, using pooled sample expression profiling J Natl Cancer Inst 2002 Apr 3,94(7) 513-21

Birkenkamp-Demtroder K, Christensen LL, Olesen SH, Frederiksen CM, Laiho P, Aaltonen LA, Laurberg S, Sorensen FB, Hagemann R, ORntoft TF Gene expression in colorectal cancer Cancer Res 2002 Aug 1,62(15) 4352-63

Boland CR, Thibodeau SN, Hamilton SR, Sidransky D, Eshleman JR, Burt RW, Meltzer SJ, Rodriguez-Bigas MA, Fodde R, Ranzani GN, Srivastava S A National Cancer Institute Workshop on Microsatellite Instability for cancer detection and familial predisposition development of international criteria for the determination of microsatellite instability in colorectal cancer Cancer Res 1998 Nov 15,58(22) 5248-57 Review

Chapusot C, Martin L, Bouvier AM, Bonithon-Kopp C, Ecarnot-Laubriet A, Rageot D, Ponnelle T, Laurent Puig P, Faivre J, Piard F Microsatellite instability and intratumoural heterogeneity in 100 right-sided sporadic colon carcinomas Br J Cancer 2002 Aug 12,87(4)·400-4

Dyrskjot L, Thykjaer T, Kruhoffer M, Jensen JL, Marcussen N, Hamilton-Dutoit S, Wolf H, Orntoft TF Identifying distinct classes of bladder carcinoma using microarrays Nat Genet 2003 Jan,33(1) 90-6

Frederiksen CM, Knudsen S, Laurberg S, Orntoft TF Classification of Dukes' B and C colorectal cancers using expression arrays J Cancer Res Clin Oncol. 2003 May,129(5).263-71.

Huang J, Qi R, Quackenbush J, Dauway E, Lazaridis E, Yeatman T Effects of ischemia on gene expression J Surg Res 2001 Aug;99(2)·222-7

Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP Summaries of Affymetrix GeneChip probe level data Nucleic Acids Res. 2003 Feb 15;31(4) e15

Loukola A, Eklin K, Laiho P, Salovaara R, Kristo P, Jarvinen H, Mecklin JP, Launonen V, Aaltonen LA Microsatellite marker analysis in screening for hereditary nonpolyposis colorectal cancer (HNPCC) Cancer Res 2001 Jun 1,61(11) 4545-9

Markowitz S, Hines JD, Lutterbaugh J, Myeroff L, Mackay W, Gordon N, Rustum Y, Luna E, Kleinerman J Mutant K-ras oncogenes in colon cancers Do not predict Patient's chemotherapy response or survival Clin Cancer Res 1995 Apr;1(4) 441-5.

Mori Y, Selaru FM, Sato F, Yin J, Simms LA, Xu Y, Olaru A, Deacu E, Wang S, Taylor JM, Young J, Leggett B, Jass JR, Abraham JM, Shibata D, Meltzer SJ. The impact of microsatellite instability on the molecular phenotype of colorectal tumors Cancer Res. 2003 Aug 1,63(15) 4577-82

Ribic CM, Sargent DJ, Moore MJ, Thibodeau SN, French AJ, Goldberg RM, Hamilton SR, Laurent-Puig P, Gryfe R, Shepherd LE, Tu D, Redston M, Gallinger S Tumor microsatellite-instability status as a predictor of benefit from fluorouracil-based adjuvant chemotherapy for colon cancer
N Engl J Med 2003 Jul 17,349(3) 247-57

Figures

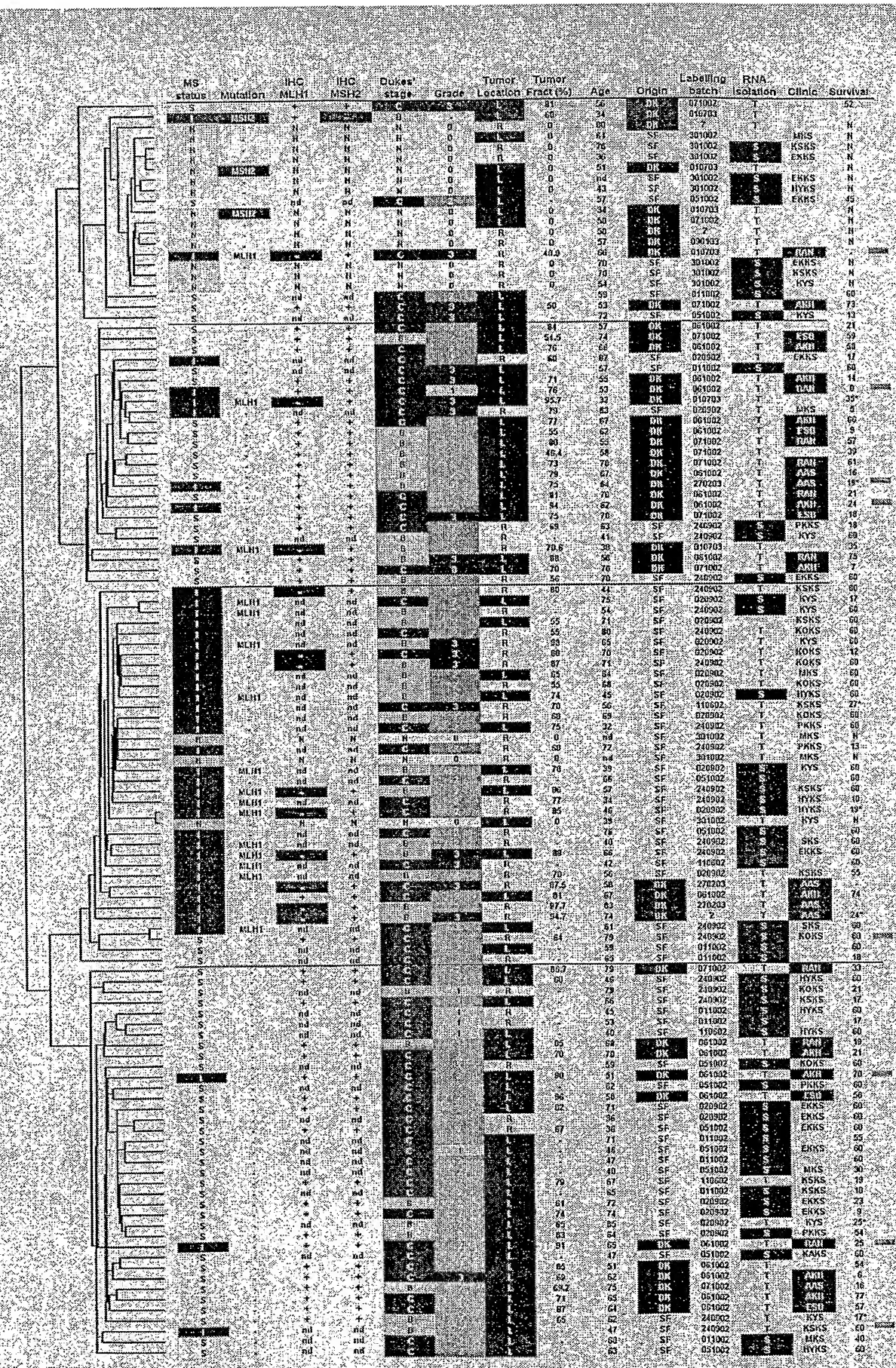


Figure 1. Clusteranalysis of Colon Specimens with Associated Clinicopathological Features.

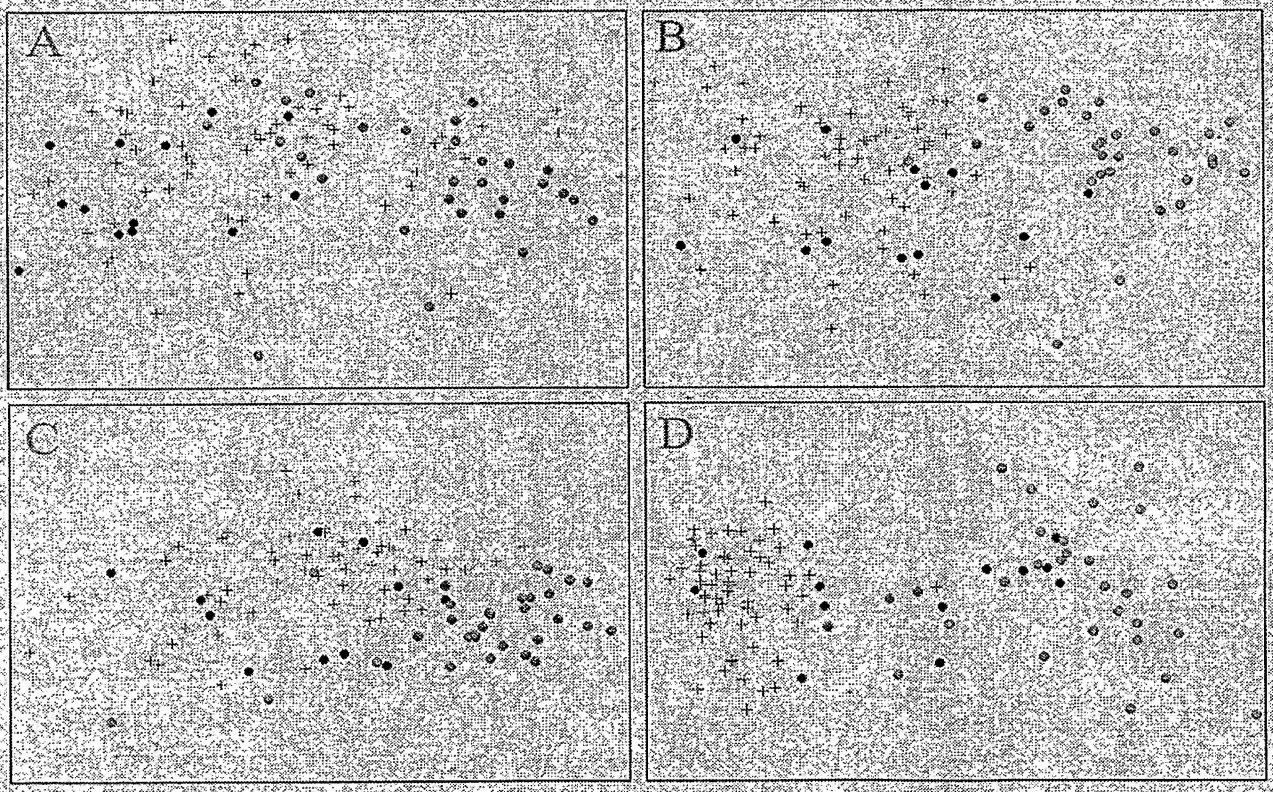
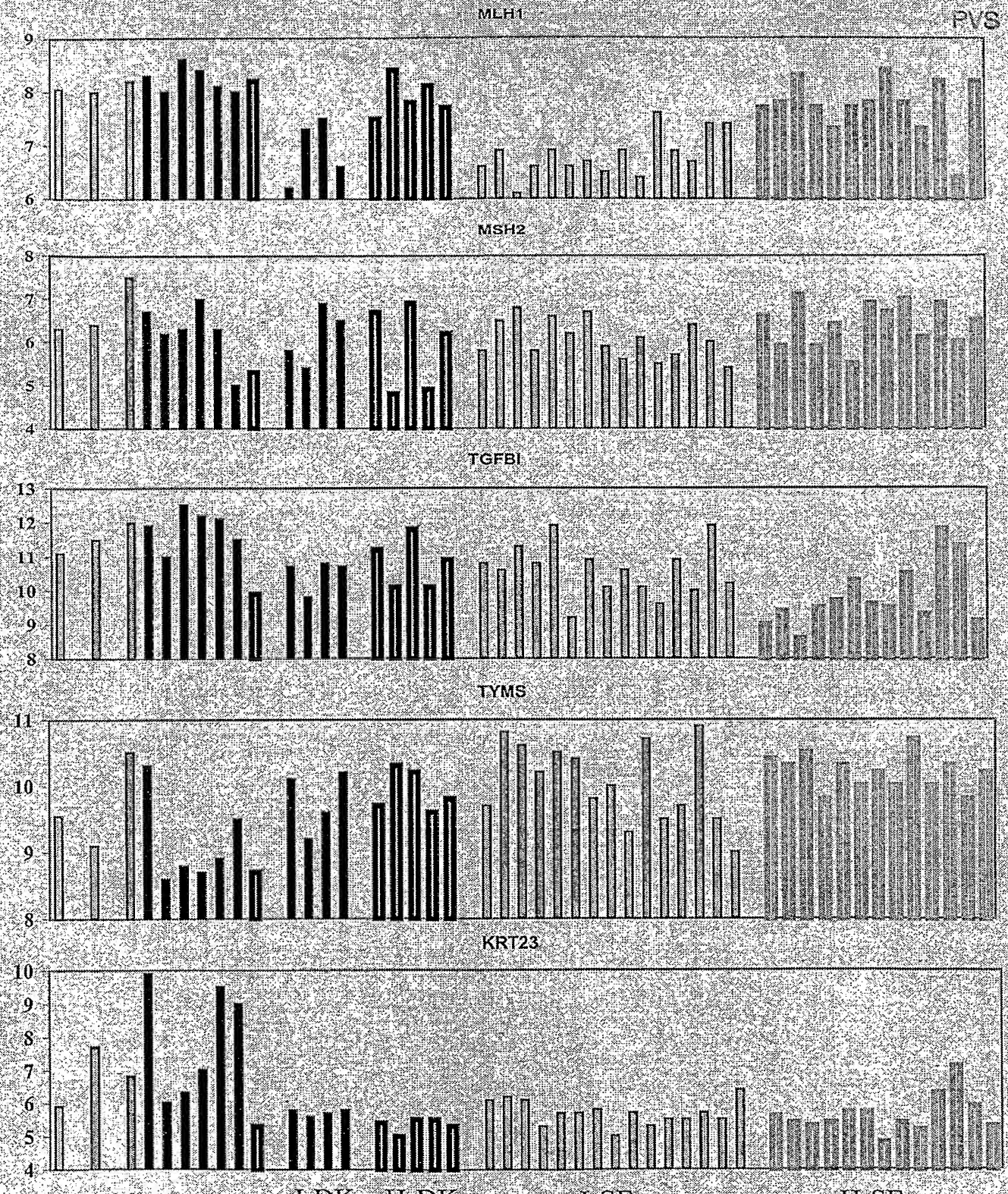


Figure 2. Multidimentional Analysis showing distances between groups of tumors.



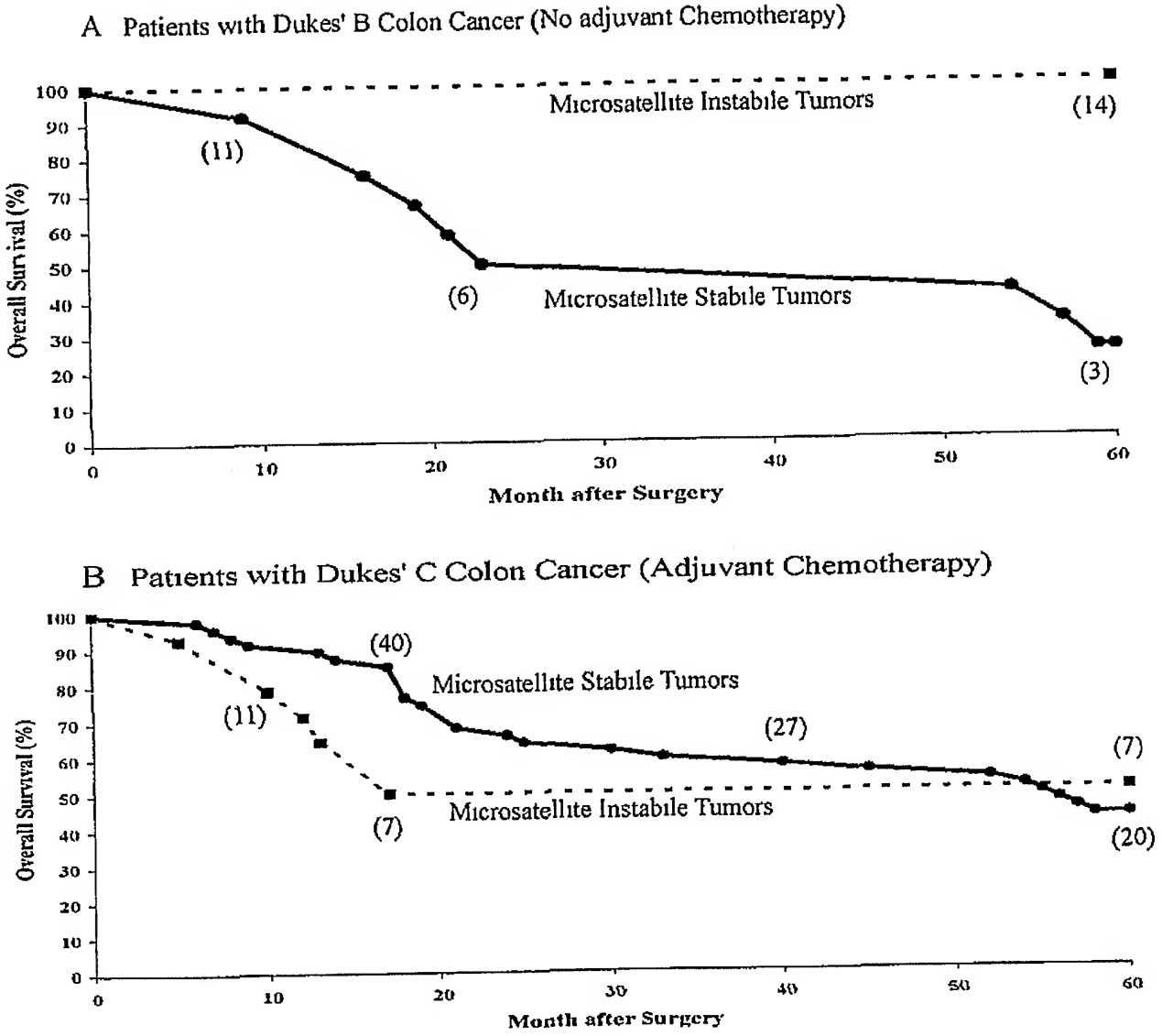


Figure 4. Kaplan-Meier Estimates of Overall Survival among Patients with Dukes' B and Dukes' C Colon Cancer According to the Microsatellite-Instability Status of the Tumor.

Table 1. Summary of clinocopathological and microsatellite features of colon samples

Patient group n (DK,SF)	Median age range	Localization in colon right (DK,SF)	left (DK,SF)	Dukes' Stage n (DK,SF) N[a]	B	C	IHC negative stain N (n tested) MLH1	MSH2
All cases 119 (44,75)	62 0	45 (8,37)	74 (36,38)	17 (6,11)	36 (14,22)	66 (20,46)	12 (56)	1 (56)
MSI-H[b] 24 (9,16)	67 0	15 (3,12)	9 (6,4)	-	10 (3,7)	14 (5,9)	6 (11)	0 (11)
HNPCC[c] 17(4,13)	45 0	9 (2,7)	8 (2,6)	-	10 (2,8)	7 (2,5)	6 (8)	1 (8)
MSS 60 (25,35)	63 0	11 (0,11)	49 (25,24)	-	16 (9,7)	44 (16,28)	0 (37)	0 (37)

[a]normal biopsy taken from the resection edge of a tumor

[b]sporadic tumors

[c]all tumors MSI-H

Table 2 Permutation test of groups

Pseudo group	S1/S2 from data	Smaller values in 100 permutations	Minimum in 100 permutations
DK-SF	0.9072795	0	0.962269
I-S	0.9166195	0	0.9583325

Table 3. Permutation test of genes

		$S_1(j)/S_2(j)$			
Pseudo group		< 0.6	< 0.7	< 0.8	< 0.9
DK-SF	number of genes	36	136	522	1785
	max in 100 permutations	0	0	2	225
MSI-MSS	number of genes	17	103	399	1507
	max in 100 permutations	0	1	8	250

Table 4. Genes for classification of MSS and MSI tumors

Probe_ID	Symbol	Unigene	Name
211456_x_at	AF333388	AF333388	AF333388
216336_x_at	AL031602	AL031602	AL031602
217917_s_at	DNCL2A	Hs 100002	dynein, cytoplasmic, light polypeptide 2A
205242_at	CXCL13	Hs 100431	chemokine (C-X-C motif) ligand 13 (B-cell chemoattractant)
203315_at	NCK2	Hs 101695	NCK adaptor protein 2
205495_s_at	GNLY	Hs 105806	granulysin
209546_s_at	APOL1	Hs 114309	apolipoprotein L, 1
201641_at	BST2	Hs 118110	bone marrow stromal cell antigen 2
212185_x_at	MT2A	Hs 118786	metallothionein 2A
209108_at	TM4SF6	Hs 121068	transmembrane 4 superfamily member 6
218345_at	HCA112	Hs 12126	hepatocellular carcinoma-associated antigen 112
210538_s_at	BIRC3	Hs 127799	baculoviral IAP repeat-containing 3
206754_s_at	CYP2B6	Hs 1360	cytochrome P450, family 2, subfamily B, polypeptide 6
218346_s_at	PA26	Hs 14125	p53 regulated PA26 nuclear protein
217759_at	TRIM44	Hs 14512	tripartite motif-containing 44
218943_s_at	RIG-I	Hs 145612	DEAD/H (Asp-Glu-Ala-Asp/His) box polypeptide
219956_at	GALNT6	Hs 151678	UDP-N-acetyl-alpha-D-galactosamine polypeptide N-acetylgalactosaminyltransferase 6 (GalNAc-T6)
206907_at	TNFSF9	Hs 1524	tumor necrosis factor (ligand) superfamily, member 9
203008_x_at	APACD	Hs 153884	ATP binding protein associated with cell differentiation
218898_at	CT120	Hs 154396	membrane protein expressed in epithelial-like lung adenocarcinoma
218325_s_at	DATF1	Hs 155313	death associated transcription factor 1
206918_s_at	CPNE1	Hs 166887	copine I
203444_s_at	MTA1L1	Hs 173043	metastasis-associated 1-like 1
204070_at	RARRES3	Hs 17466	retinoic acid receptor responder (tazarotene induced) 3
212349_at	POFUT1	Hs 178292	protein O-fucosyltransferase 1
218237_s_at	SLC38A1	Hs 18272	solute carrier family 38, member 1
201910_at	FARP1	Hs 183738	FERM, RhoGEF (ARHGEF) and pleckstrin domain protein 1 (chondrocyte-derived)
212229_s_at	FBXO21	Hs 184227	F-box only protein 21
218704_at	FLJ20315	Hs 18457	hypothetical protein FLJ20315
208156_x_at	EPPK1	Hs 200412	epiplakin 1
214617_at	PRF1	Hs 2200	perforin 1 (pore forming protein)
201884_at	CEACAM5	Hs 220529	carcinoembryonic antigen-related cell adhesion molecule 5
208022_s_at	CDC14B	Hs 22116	CDC14 cell division cycle 14 homolog B (S. cerevisiae)
220658_s_at	ARNTL2	Hs 222024	transcription factor BMAL2
204533_at	CXCL10	Hs 2248	chemokine (C-X-C motif) ligand 10
218802_at	FLJ20647	Hs 234149	hypothetical protein FLJ20647
221653_x_at	APOL2	Hs 241412	apolipoprotein L, 2
207457_s_at	LY6G6D	Hs 241587	lymphocyte antigen 6 complex, locus G6D
213470_s_at	HNRPH1	Hs 245710	heterogeneous nuclear ribonucleoprotein H1 (H)
202262_x_at	DDAH2	Hs 247362	dimethylarginine dimethylaminohydrolase 2
218094_s_at	C20orf35	Hs 256086	chromosome 20 open reading frame 35
217727_x_at	VPS35	Hs 264190	vacuolar protein sorting 35 (yeast)
204415_at	G1P3	Hs 265827	interferon, alpha-inducible protein (clone IFI-6-16)
208461_x_at	MT1H	Hs 2667	metallothionein 1H
202072_at	HNRPL	Hs 2730	heterogeneous nuclear ribonucleoprotein L
205241_at	SCO2	Hs 278431	SCO cytochrome oxidase deficient homolog 2 (yeast)
203896_s_at	PLCB4	Hs 283006	phospholipase C, beta 4
221920_s_at	MSCP	Hs 283716	mitochondrial solute carrier protein
213385_at	CHN2	Hs 286055	chimerin (chimaerin) 2
204020_at	PURA	Hs 29117	purine-rich element binding protein A
202589_at	TYMS	Hs 29475	thymidylate synthetase
210321_at	CTLA1	Hs 348264	similar to granzyme B (granzyme 2, cytotoxic T-lymphocyte-associated serine esterase 1) (H. sapiens)
206976_s_at	HSPH1	Hs 36927	heat shock 105kDa/110kDa protein 1
208581_x_at	MT1X	Hs 374950	metallothionein 1X
204326_x_at	MT1L	Hs 380778	metallothionein 1L
209504_s_at	PLEKHB1	Hs 380812	pleckstrin homology domain containing, family B (evectins) member 1
204131_s_at	FOXO3A	Hs 380831	forkhead box O3A
215780_s_at	Hs 382039	Hs 382039	Homo sapiens, clone IMAGE 4420333, mRNA
212341_at	Hs 405983	Hs 405983	Homo sapiens cDNA FLJ21020 fis, clone CAE06067
213738_s_at	ATP5A1	Hs 405985	ATP synthase, H+ transporting, mitochondrial F1 complex, alpha subunit, isoform 1, cardiac muscle
207993_s_at	CHP	Hs 406234	calcium binding protein P22
201849_at	UBE2L6	Hs 425777	ubiquitin-conjugating enzyme E2L 6
212859_x_at	MT1E	Hs 433205	metallothionein 1E (functional)
204745_x_at	MT1G	Hs 433391	metallothionein 1G
202520_s_at	MLH1	Hs 433518	mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli)
201762_s_at	PSME2	Hs 433810	proteasome (prosome, macropain) activator subunit 2 (PA28 beta)
218242_s_at	CGI-85	Hs 442630	CGI-85 protein
222244_s_at	FLJ20618	Hs 52184	hypothetical protein FLJ20618
202762_at	ROCK2	Hs 58617	Rho-associated, coiled-coil containing protein kinase 2
207320_x_at	STAU	Hs 6113	staufen, RNA binding protein (Drosophila)
201976_s_at	MYO10	Hs 61638	myosin X
215693_x_at	DDX27	Hs 65234	DEAD/H (Asp-Glu-Ala-Asp/His) box polypeptide 27
212070_at	GPR56	Hs 6527	G protein-coupled receptor 56
214924_s_at	OIP106	Hs 6705	OGT(O-Glc-NAc transferase)-interacting protein 106 KDa
206108_s_at	SFRS6	Hs 6891	splicing factor, arginine/serine-rich 6

Table

204858_s_at	ECGF1	Hs 73946	endothelial cell growth factor 1 (platelet-derived)
213201_s_at	TNNT1	Hs 73980	troponin T1, skeletal, slow
200814_at	PSME1	Hs 75348	proteasome (prosome, macropain) activator subunit 1 (PA28 alpha)
206286_s_at	TDGF1	Hs 75561	teratocarcinoma-derived growth factor 1
204103_at	CCL4	Hs 75703	chemokine (C-C motif) ligand 4
203559_s_at	ABP1	Hs 75741	amiloride binding protein 1 (amine oxidase (copper-containing))
209048_s_at	PRKCBP1	Hs 75871	protein kinase C binding protein 1
202678_at	GTF2A2	Hs 76362	general transcription factor IIA, 2, 12kDa
203915_at	CXCL9	Hs 77367	chemokine (C-X-C motif) ligand 9
201674_s_at	AKAP1	Hs 78921	A kinase (PRKA) anchor protein 1
202203_s_at	AMFR	Hs 80731	autocrine motility factor receptor
203773_x_at	BLVRA	Hs 81029	biliverdin reductase A
208944_at	TGFBR2	Hs 82028	transforming growth factor, beta receptor II (70/80kDa)
200628_s_at	WARS	Hs 82030	tryptophanyl-tRNA synthetase
204780_s_at	TNFRSF6	Hs 82359	tumor necrosis factor receptor superfamily, member 6
220951_s_at	ACF	Hs 8349	apobec-1 complementation factor
221516_s_at	FLJ20232	Hs 83869	hypothetical protein FLJ20232
217875_s_at	TMEPAI	Hs 83883	transmembrane, prostate androgen induced RNA
210029_at	INDO	Hs 840	indoleamine-pyrrole 2,3 dioxygenase
204044_at	QPRT	Hs 8935	quinolinate phosphonbosyltransferase (nicotinate-nucleotide pyrophosphorylase (carboxylating))
218963_s_at	KRT23	Hs 9029	keratin 23 (histone deacetylase inducible)

Table 5 Performance of the classifier

	Trainings set Errors in crossvalidation	Test set Test errors
MSI	7.2% (n=25, range 0-4)	6.8% (n=10, range 0-3)
MSS	0.13% (n=30, range 0-1)	0.17% (n=29, range 0-3)
All	3.8% (n=55, range 0-4)	1.8% (n=39, range 0-3)